Demystifying Scientific Problem-Solving in LLMs by Probing Knowledge and Reasoning

Alan Li[1*] **Yixin Liu**[1*] **Arpan Sarkar**[2]
Doug Downey[3,4] **Arman Cohan**[1,4]

[1]Yale University, [2]Harvard University, [3]Northwestern University, [4]Allen Institute of AI
{haoxin.li,yixin.liu}@yale.edu

Abstract

Scientific problem solving poses unique challenges for LLMs, requiring both deep domain knowledge and the ability to apply such knowledge through complex reasoning. While automated scientific reasoners hold great promise for assisting human scientists, there is currently no widely adopted holistic benchmark for evaluating scientific reasoning, and few approaches systematically disentangle the distinct roles of knowledge and reasoning in these tasks. To address these gaps, we introduce SCIREAS, a diverse suite of existing benchmarks for scientific reasoning tasks, and SCIREAS-PRO, a selective subset that requires more complex reasoning. Our holistic evaluation surfaces insights about scientific reasoning performance that remain hidden when relying on individual benchmarks alone. We then propose KRUX, a probing framework for studying the distinct roles of reasoning and knowledge in scientific tasks. Combining the two, we conduct an in-depth analysis that yields several key findings: (1) Retrieving task-relevant knowledge from model parameters is a critical bottleneck for LLMs in scientific reasoning; (2) Reasoning models consistently benefit from external knowledge added in-context on top of the reasoning enhancement; (3) Enhancing verbalized reasoning improves LLMs' ability to surface task-relevant knowledge.[1]

1 Introduction

Recent frontier reasoning models, such as OpenAI's o-series [33] and DeepSeek-R1 [8], demonstrate significant advances by leveraging increased test-time compute to enable intermediate reasoning steps [54, 19]. These approaches facilitate advanced mechanisms, including methodology exploration [59], self-verification [29], and backtracking [58], resulting in improvements on tasks such as mathematics and coding with more test-time compute [32].

These advances in reasoning capabilities create opportunities for applying LLMs to complex scientific tasks [27, 13, 38]. However, scientific work demands not only rigorous reasoning but also deep domain knowledge, from specialized concepts and foundational theories to hands-on methodological expertise and familiarity with obscure yet pivotal findings. Successful scientific reasoning systems must apply such knowledge in complex multi-step reasoning processes [61, 49, 46, 25].

While a variety of scientific benchmarks exist (e.g., GPQA [37] and MMLU-Pro [52]), there is no holistic and unified benchmark that comprehensively targets scientific reasoning. Existing individual benchmarks typically focus narrowly on specific domains, task formats, or skill types. For example, although GPQA is challenging, it focuses exclusively on multiple-choice questions within a limited

[*]These authors contributed equally to this work.

[1]The codebase and artifacts are released at https://github.com/yale-nlp/SciReas-Eval.

Preprint. Under review.

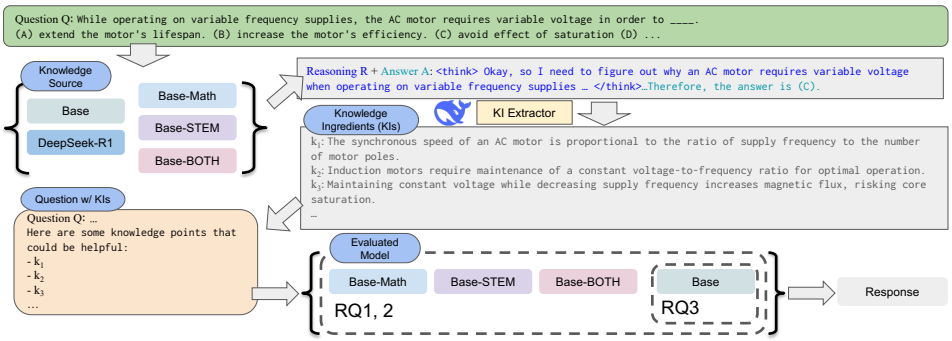

Figure 1: KRUX pipeline. Starting from the upper left, we prompt an LLM (one of base, DeepSeek-R1, Base-Math, Base-STEM, and Base-BOTH) with a question from SCIREAS as knowledge source, collect the output and reasoning traces, and feed the reasoning traces to DeepSeek-R1 as the extractor to generate knowledge ingredients (KIs). We then evaluate the tested model with KI-augmented questions, which allows us to study three key research questions (RQ1, RQ2, RQ3) regarding LLMs' knowledge and reasoning capabilities in scientific problem-solving.

range of domains. Furthermore, there is a lack of analytical tools that can isolate the distinct roles that reasoning and scientific knowledge play when performing sophisticated scientific tasks.

We introduce datasets and methods to facilitate the study of scientific problem solving. First, we present **SCIREAS**, a unified suite of ten public benchmarks that span physics, chemistry, biology, medicine, materials, mathematics, computer science, and engineering, with multiple-choice, fill-in-the-blank, structured, and protocol/procedural questions. To improve evaluation efficiency and sharpen the focus on reasoning difficulty, we manually inspect each subtask and retain only those that are subject-relevant and reasoning-intensive, while preserving broad domain coverage. Furthermore, to facilitate standardized evaluation, we provide an efficient and unified implementation of streamlined assessment across individual benchmarks, avoiding the need to set up different environments or dataset-specific boilerplate for each dataset (§3).

Next, we introduce **SCIREAS-PRO**, a compact subset of SCIREAS tailored for evaluating more challenging reasoning. Specifically, SCIREAS-PRO is constructed by selecting examples from SCIREAS where only reasoning models with high inference-time compute budget (or the highest allowed number of thinking tokens) succeed. We find that despite containing only 8% as many examples as SCIREAS, SCIREAS-PRO better differentiates weak and strong reasoners (§3).

Having constructed the reasoning-intensive scientific benchmarks, our next goal is to leverage them to study how verbalized chain-of-thought (CoT) reasoning affects knowledge recall and usage (§4). To study this, we design **KRUX** (Knowledge & Reasoning Utilization eXams), a probing framework which supplies models with atomic "knowledge ingredients" (KIs) extracted from other models' reasoning traces (Figure 1). This technique allows for more controlled analyses of reasoning and knowledge, which we use to perform three in-depth investigations that lead to the following findings:

(1) Vanilla instruct models can *outperform* their reasoning counterparts by $\geq 10\%$ once KIs are provided in-context, **suggesting that internalizing and retrieving the right knowledge is a key bottleneck for scientific reasoning tasks.**

(2) When both model families receive the same KIs from a strong reasoner (e.g., DeepSeek-R1), the reasoning-fine-tuned models consistently outperform the base models, showing that **reasoning models are capable of utilizing external in-context knowledge for additional improvements.**

(3) Feeding KIs from a reasoning-fine-tuned model to its base model can boost performance even when the KIs are already known by the base model, indicating that **reasoning-fine-tuning aids knowledge recall by surfacing more relevant knowledge**.

Our contributions can be summarized as:

- We introduce SCIREAS, a unified and holistic benchmark suite spanning a broad range of scientific domains and problem types, allowing us to surface insights that otherwise remain hidden if relying on individual datasets only. We also release a reasoning-focused subset SCIREAS-PRO that allows efficient benchmarking of sophisticated reasoning with more room for improvement.

• We present KRUX, a novel analytic framework which we use to conduct a comprehensive empirical study that disentangles the impacts of knowledge and reasoning.
• We provide an in-depth analysis with three key findings: (i) knowledge retrieval is a bottleneck; (ii) in-context knowledge consistently benefits reasoning models; and (iii) long CoT improves knowledge surfacing. We support these findings with controlled post-training experiments, and show our training recipe is competitive compared with concurrent SFT post-training efforts.

2 Related Work

Scientific Benchmarks Existing scientific benchmarks span a wide array of domains and tasks, but each tends to focus on specific disciplines or subskills, often lacking explicit emphasis on multi-step reasoning or standardized implementation. For example, most tasks in SciRIFF [46] focus on context-grounded information QA, rather than demanding reasoning. Benchmarks like GPQA [37] and LabBench [22] pose reasoning challenges, yet they cover only a limited range of scientific domains and rely on multiple-choice QA formats. Implementation-wise, benchmarks lack standardized prompts, evaluation metrics, or consistent scoring, making reproducibility and fair comparison difficult [15, 11].

To address this fragmentation, our study systematically incorporates 10 prominent scientific benchmarks, GPQA, MMLU-Pro [52], SuperGPQA [41], LabBench, OlympiadBench [16], SciBench [51], SciRIFF, UGPhysics [56], SciEval [39], and SciKnowEval [10], enabling a unified, comprehensive, and reproducible evaluation of scientific reasoning capabilities.

Knowledge & Reasoning An important line of work on disentangling reasoning and knowledge designs specialized tasks (e.g., linguistically challenging questions [2, 18] or synthetic multi-hop questions [24]) to isolate reasoning from knowledge, but such benchmarks are often artificial and domain-constrained. Notably, Li and Goyal [24] analyzes the synergy between knowledge and reasoning as knowledge evolves, offering a perspective complementary to our controlled CoT SFT experiments. Another line of work trains external classifiers to label questions as reasoning- or knowledge-intensive based on parametric models [44]. However, this approach requires well-calibrated training data and does not distinguish the tested model's internal knowledge from reasoning; for instance, a question labeled as requiring reasoning might be directly memorized by the model. Concurrent work leverages reasoning traces to evaluate factual correctness [55], but focuses on surface-level factuality rather than genuine knowledge recall. Unlike prior work that trains external classifiers to label question types or checks surface factuality in traces, KRUX holds knowledge constant and varies the target model, isolating knowledge recall from reasoning ability without relying on heuristic difficulty tags. Additional related work is provided in Appendix B.

3 Benchmarking Knowledge-Intensive Scientific Reasoning

Given limited coverage in terms of domain, formats, or accessibility for individual benchmarks, SCIREAS solves this by merging ten datasets under one standardized harness, offering broad domain coverage and consistent evaluation.

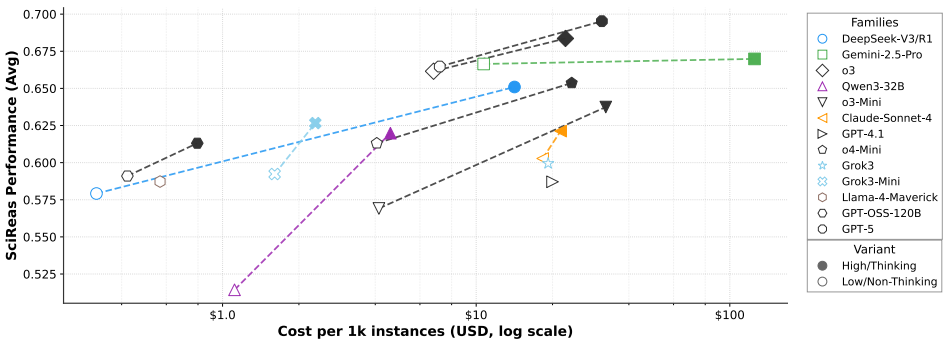

Figure 2: Frontier reasoning models' performance evaluated on SCIREAS. The X-axis shows the cost per 1k instances in USD. Different reasoning settings on the same model can result in distinct costs and performance, but the margins vary depending on the models.

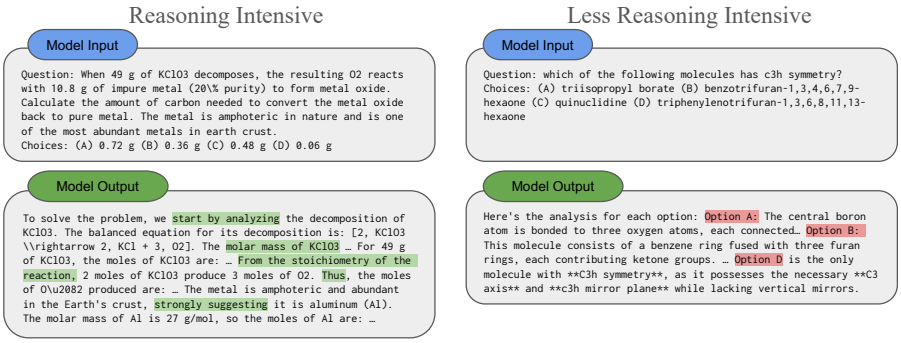

Figure 3: An example pair with varying reasoning intensity, where the example on the left is sampled from SCIREAS-PRO and the right is a filtered out example (§3). On the left, the progressive reasoning chain is highlighted. The example on the right emphasizes knowledge recall on each option with a simple elimination strategy.

SCIREAS SCIREAS is a unified evaluation suite focused on reasoning-intensive scientific tasks curated from 10 representative existing benchmarks. Through task-level filtering, SCIREAS reduces instance count by nearly 50% while preserving coverage, and, inspired by OLMES [15], provides a unified implementation optimized with vLLM [20] and batch job APIs[2] for scalable, easy-to-use, and efficient evaluation. Figure 2 shows the performance of various LLMs on SCIREAS (more in §3.1).

Our curation prioritizes subtasks from each benchmark that demand not only specific domain knowledge but also complex, multi-step reasoning processes for resolution. For each subtask from each benchmark, we manually inspect at least 20 instances. We manually determine (1) whether the given task requires an in-depth understanding of domain-specific scientific knowledge beyond the information provided in-context, and (2) whether multi-step reasoning is necessary to reach the correct answer. We incorporate tasks only if *all* 20 sampled instances fulfill both requirements.[3]

To keep evaluation cost-efficient, we uniformly sample 200 instances from each subtask sourced from high-cost benchmarks — MMLU-Pro, SciKnowEval, SciEval, and UGPhysics, which maintains similar evaluation outcomes (more in Appendix C.2) while reducing the cost by nearly 50% (from 29,604 to 15,567 total instances). The complete list of selected subtasks, their subject coverage, data sources, and evaluation metrics are in Appendix C.1. Benchmarks affected by our filtering are marked with an asterisk (*); their scores are not directly comparable to those from prior work.

SCIREAS-PRO Although SCIREAS provides a uniform measurement for model performance on scientific reasoning tasks that nominally require scientific reasoning, the difficulty of individual instances is uneven: some can be answered with little or no deductive effort once the pertinent fact is recalled, as shown in an example in Figure 3.

To isolate the reasoning skill, we therefore curate a "hard" subset — those questions whose solutions still demand multi-step inference even when all relevant knowledge is available — so that any performance gains cannot be explained by knowledge recall alone. Building on our observation in §3.1.1, we hypothesize that the performance difference under different test-time inference budgets can serve as an effective indicator of reasoning intensity. Specifically, instances where reasoning models fail with low reasoning budget but succeed with high budget likely require complex reasoning processes, even when the necessary domain knowledge is accessible to the model in both settings.

In practice, we evaluate o3-mini and o4-mini on SCIREAS with both *high* and *low* "reasoning-effort" settings — an OpenAI API flag that limits the number of thinking tokens before the answer. For o3-mini and o4-mini, the high-effort setting costs at least 5.8× more per instance than the low-effort setting (Table 6, Appendix C.1).[4] For each model, we keep questions answered incorrectly under low effort but correctly under high effort and take the union of these sets to create SCIREAS-PRO,

[2]We provide batch job inference options for popular LLM providers, e.g., OpenAI, Anthropic, TogetherAI, and Gemini. Using batch APIs allows for up to 50% cost reduction.

[3]While this manual inspection can be subjective, it is based on the authors' graduate-school-level expertise.

[4]Because these models are proprietary, factors beyond the flag may influence performance. We therefore treat the flag as a practical, not absolute, proxy and validate it with an independent LLM-judge study (Appendix C.3.2).

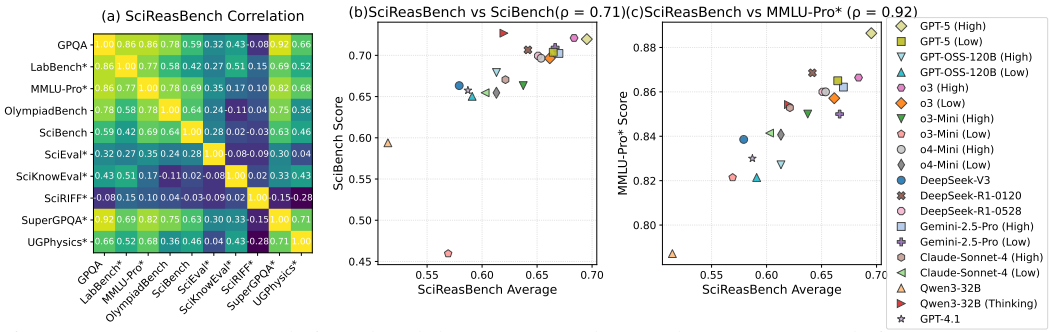

Figure 4: SCIREAS correlations breakdown. (a) Task-to-task Pearson correlations. SCIREAS incorporates tasks complementary to popular benchmarks. (b) and (c) show performance on SCIREAS vs. SciBench and MMLU-Pro*. Models may be tuned for certain tasks, outperforming higher-ranked models on individual benchmarks.

resulting in 1,260 unique instances. We further validate this approach by using LLM judge as well as human evaluation to check the reasoning-intensiveness of resulting examples from this filtering pipeline. Appendix C.3 shows that both human annotators and LLM judges find SCIREAS-PRO to be indeed richer in reasoning-intensive instances.

3.1 Benchmarking Frontier Models

Having constructed SCIREAS and SCIREAS-PRO with focus on scientific reasoning tasks, we now examine how state-of-the-art models perform under varying computational budgets. We evaluate frontier models using different "reasoning-effort" settings (see configuration details in Appendix D). These settings typically correspond to significant differences in output length, with high-effort modes producing substantially more reasoning tokens as they work through complex problems.[5]

3.1.1 Results

Aggregated Results Figure 2 highlights aggregated performance evaluated on SCIREAS, with score breakdowns on selected models shown in Table 6. Notably, the **aggregated ranking provides additional insights that differ from popular individual benchmarks**. Comparing o3-High and Gemini-2.5-Pro-Preview-High as an example, o3-High wins on GPQA and MMLU-Pro* while Gemini-2.5-Pro-Preview-High wins on SuperGPQA*, all with a thin margin (within 1 absolute point, even evaluated on MMLU-Pro before uniform sampling as shown in Figure 7). Similarly, GPT-5-High shows on-par performance with Gemini-2.5-Pro-Preview-High on problem-solving benchmarks like OlympiadBench and SciRIFF. Evaluated across SCIREAS, however, we notice that GPT-5-High outperforms its competitors on a broader range of benchmarks. Meanwhile, o3-High achieves higher overall performance over Gemini-2.5-Pro-Preview-High, with superior performance on LabBench* and weaker on OlympiadBench by a large margin (beyond 10 absolute points).

Benchmark Correlations In general, as the Pearson correlation analysis shows in Figure 4 (a), while some benchmarks are closely correlated (e.g., GPQA and SuperGPQA*), benchmarks containing free-form QA and fill-in-the-blank questions like SciRIFF* and SciEval* are not highly correlated with GPQA-like multiple-choice tasks, demonstrating the need for a holistic evaluation suite.

Isolating specific benchmarks, we observe that **models from different providers may be tuned explicitly for specific tasks or skills**. As shown in Figure 4 (b) and (c), Qwen3-32B-Thinking strikes noticeably above the trend on SciBench, reaching comparable performance to commercial frontier models. Similarly, DeepSeek-V3 and DeepSeek-R1-0120 demonstrate stronger performance on MMLU-Pro*, indicating capabilities that surpass their overall rankings.

Performance Gap by Reasoning Difference Although the gap varies depending on different model families, **the same model can exhibit a significant performance gap under different reasoning settings**. For instance, as shown in Figure 2, o3-mini-Low and -High show a performance gap of

[5]In this work, we refer to DeepSeek-R1-0528 and DeepSeek-V3-0324 simply as DeepSeek-R1 and DeepSeek-V3, respectively, unless otherwise specified.

6.8. Similar traits can be observed among o4-mini, Claude-Sonnet-4, and o3, while Gemini-2.5-Pro shows the least performance gain, even with significantly more (>10×) thinking budget.

This observation motivates the construction of SCIREAS-PRO, leveraging the performance gap between low and high reasoning efforts as an effective proxy for identifying instances that demand complex reasoning rather than mere knowledge recall. **For practitioners, task-specific evaluation is still recommended** for the optimal balance between inference cost and performance.

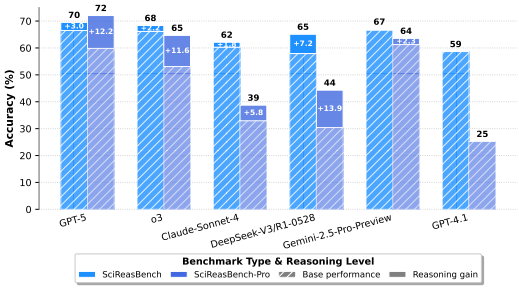

Figure 5: Model performance on SCIREAS and SCIREAS-PRO with varying reasoning capabilities. SCIREAS-PRO amplifies gaps between low-/non-reasoning and high-reasoning settings.

Amplified Performance Gap Figure 5 shows that **SCIREAS-PRO amplifies performance gaps between low- and high-reasoning settings**, where the gap between GPT-5-High and GPT-5-Low widens from 3.01 to 12.22, and the corresponding gap for Gemini-2.5-Pro-Preview widens from 0.35 to 2.30. Meanwhile, non-reasoning models, e.g., GPT-4.1, DeepSeek-V3, show more significant gaps compared to concurrent reasoning models, i.e., o3 and DeepSeek-R1, respectively.

4 Disentangling Knowledge and Reasoning in Scientific Tasks

While SCIREAS and SCIREAS-PRO provide unified benchmarks to evaluate scientific reasoning capabilities, another fundamental question remains: how does CoT reasoning adaptation affect a model's ability to recall and utilize knowledge? To address this question, we first conduct a series of controlled SFT experiments on high-quality reasoning traces with and without in-domain scientific knowledge, and then we propose KRUX, a novel investigative framework to study three key research questions regarding the role of knowledge in scientific reasoning using the fine-tuned checkpoints.

4.1 Controlled CoT SFT

To control for data composition and isolate the impact of reasoning and knowledge injection during post-training, we fine-tune Qwen2.5-7B-Instruct [57] and Llama-3.1-8B-Instruct [14] on reasoning traces drawn from mathematics and STEM domains, as well as on their combination. This allows us to attribute behavior changes to the data mixture rather than confounding factors.

Table 1: Performance of reasoning models trained from Qwen2.5-Instruct and Llama-3.1-Instruct on SYNTHETIC-1 and concurrent reasoning models.

Model	Method	SCIREAS	-PRO
Our Checkpoints			
Qwen	–	37.07	13.97
Qwen-STEM	SFT	40.47	16.11
Qwen-Math	SFT	41.99	18.17
Qwen-BOTH	SFT	**42.84**	**21.11**
Llama	–	31.25	11.67
Llama-STEM	SFT	35.28	14.29
Llama-Math	SFT	35.49	16.98
Llama-BOTH	SFT	38.55	16.51
Concurrent Reasoning Post-training			
SYNTHETIC-1-SFT	SFT	37.64	19.44
OpenR1	SFT	43.08	**26.43**
Llama-Nemotron	SFT&RL	**43.53**	23.75
General-Reasoner	RL	34.99	13.73

For training, we leverage the SYNTHETIC-1 [31] dataset, an existing large-scale dataset released by Prime Intellect,[6] which consists of outputs of DeepSeek-R1-0120, including the reasoning traces, on a diverse set of tasks. More specifically, we leverage the mathematics and STEM subsets from SYNTHETIC-1 (denoted as SYNTHETIC-1-Math/STEM, respectively). The former provides reasoning traces on abstract math reasoning questions, serving as a source for long CoT adaptation without introducing in-domain knowledge. In contrast, the latter is sourced from StackExchange [21], providing a more in-domain data source for a broader range of scientific subjects.[7] The

[6] `https://huggingface.co/PrimeIntellect`

[7] Notably, SYNTHETIC-1-Math is sourced from competition-level math problems, highlighting high-quality abstract math reasoning filtered by verified answers. In contrast, StackExchange and SYNTHETIC-1-STEM provide more realistic problem-solving data from wider subjects, offering more coverage in science domains.

math subset contains around 462K instances, while the STEM subset contains around 512K instances. Details of the training and evaluation setup are in Appendix E.

By training Qwen2.5-7B-Instruct on SYNTHETIC-1 (-Math, -STEM, and the combined subsets), we derived Qwen-Math, Qwen-STEM, and Qwen-BOTH along with their counterparts trained from Llama-3.1-8B-Instruct. In the following, we will refer to the base models as Qwen or Llama for brevity. Compared with concurrent work on long CoT post-training [4, 9, 31, 30], our checkpoints deliver comparable performance under controlled settings (Table 1), serving as reliable investigating checkpoints. A light-weight analysis on data composition is provided in Appendix E.4.

4.2 Knowledge & Reasoning Utilization Exam (KRUX)

We introduce KRUX (Figure 1), a novel investigative framework to study the role of knowledge and long CoT reasoning in scientific problem solving. To separate what a model *knows* from how it *reasons*, we hold knowledge availability fixed by injecting compact, answer-agnostic knowledge ingredients (KIs) in-context. In the framework, we extract KIs from the reasoning traces of various models and provide these KIs in-context to LLMs when evaluating them. Consequently, gains over a no-KI baseline indicate a knowledge bottleneck, while persistent errors point to reasoning limits.

We first introduce our pipeline to extract KIs from reasoning traces (§4.2.1), and then discuss how we analyze and apply extracted KIs to test knowledge recall (§4.2.2, §4.2.4) and usage (§4.2.3). For experiments, we prioritize challenging benchmarks (i.e., GPQA, MMLU-Pro*, and LabBench*), which have been widely used by previous work in the field on tasks that require scientific expertise.

4.2.1 Knowledge Ingredient (KI) Extraction

First, to analyze the role of knowledge in models' performance on scientific problem-solving, we aim to study a setting in which the model is given the requisite knowledge in-context. Specifically, we take the reasoning traces from a reasoning model as the knowledge source and use a strong reasoning-focused LLM (e.g., DeepSeek-R1) to extract the essential atomic knowledge units that comprise it, which we refer to as *knowledge ingredients* (KIs) (Figure 1). We provide the extraction prompt and example KIs in Appendix F.1. We then augment the original question by prepending the extracted KIs in-context and ask the models to solve the same problem.

We perform additional checks to ensure that KIs are relevant to the problem and do *not* leak any part of the final answer. In manual review, all extracted KIs met these criteria and were consistent with their source reasoning traces. To prevent the extractor from hallucinating or introducing extraneous facts (i.e., KIs unsupported by the source trace or unnecessary for solving the problem), we feed the generated KIs back to the source model and measure performance. If performance changes materially, this indicates potential leakage of steps or answers. Empirically, we observe no significant change (Table 2, Base vs. w/ Base KIs), suggesting the KIs are answer-agnostic and faithful to the trace. Further, although it is possible that the knowledge pieces may be irrelevant to the solution, as shown in recent studies of CoT faithfulness [45, 53, 50], recent high-performing models like DeepSeek-R1 have demonstrated strong reasoning adherence on benchmark tasks [8]. Our experiments show that the knowledge pieces help models on reasoning tasks. See Figures 12-14 in Appendix F.1 for KI examples generated by different models for the same question.

Centered on our main objective of studying knowledge recall and use in reasoning models, we examine the following key research questions: **RQ1:** To what extent can base models benefit from high-quality external knowledge? **RQ2:** Do reasoning-enhanced models benefit from external knowledge? **RQ3:** Does reasoning fine-tuning improve models' ability to surface helpful knowledge?

4.2.2 RQ1: To what extent can base models benefit from high-quality external knowledge?

Problem Statement. We investigate the potential improvement from external knowledge by providing KIs to the base models in the prompt when performing scientific reasoning (Figure 1). Here, we focus on two sources for the KIs, which are extracted from their own CoT traces (w/ Base KIs) or from DeepSeek-R1's CoT traces (w/ R1 KIs). To overcome context sensitivity, we report averages and standard deviations across 5 runs with corresponding KIs permuted randomly. We then **investigate whether there are significant gaps between base models augmented with additional KIs in the context, and their corresponding reasoning-fine-tuned models.** To this end, comparisons are made

with reasoning-fine-tuned models trained on our controlled data mixtures and the ones from concurrent work (i.e., General-Reasoner-7B [26] and Llama-Nemotron-Nano-8B [5]) that involve SFT and reinforcement learning based on the same base models (i.e., Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, respectively).

Answer to RQ1: As an upper bound, a base model with high-quality in-context knowledge can substantially outperform its reasoning-enhanced counterpart.

As shown in Table 2, base models provided with KIs from DeepSeek-R1 are able to *outperform* base models alone or Base w/ Base KIs setup by $\geq 20\%$, and *outperform* reasoning variants without KIs by $\geq 10\%$ across different benchmarks and model families, showing the external knowledge provides greater gain than reasoning fine-tuning. The fact that a base model without strong reasoning capabilities can outperform reasoning models in this setting indicates a potential deficiency of the models in knowledge recall that hinders their performance in scientific reasoning.

Table 2: Performance on GPQA and LabBench* with base models alone, base models with knowledge extracted from DeepSeek-R1 or itself (w/ {R1, Base} KIs), and base models with reasoning-fine-tuning. Best and second best average scores are labeled in bold and underlined, respectively. Reasoning models fall behind base models augmented with in-context knowledge.

Setup	GPQA	LabBench*
Qwen	35.27	32.38
w/ Qwen KIs	34.24 ± 0.93	30.93 ± 1.43
w/ R1 KIs	**47.19 ± 1.53**	**41.40 ± 2.46**
Qwen-STEM	<u>41.63</u>	31.75
Qwen-Math	39.47	30.18
Qwen-BOTH	40.81	33.83
General-Reasoner	35.94	<u>35.58</u>
Llama	28.13	33.55
w/ Llama KIs	29.06 ± 1.44	34.40 ± 2.58
w/ R1 KIs	**43.57 ± 0.88**	**42.27 ± 1.60**
Llama-STEM	38.95	36.04
Llama-Math	36.16	34.78
Llama-BOTH	<u>39.43</u>	<u>36.61</u>
Llama-Nemotron	37.95	27.78

4.2.3 RQ2: Do reasoning-enhanced models benefit from external knowledge?

Table 3: Accuracy of Qwen and Llama variants on benchmarks with external knowledge ingredients (KIs). We report averages and standard deviations over 5 random permutations of the KIs. Reasoning variants w/ R1 KIs outperform base model w/ R1 KIs across different benchmarks and models.

	GPQA		MMLU-Pro*		LabBench*	
Models	w/ self KIs	w/ R1 KIs	w/ self KIs	w/ R1 KIs	w/ self KIs	w/ R1 KIs
Qwen	34.24 ± 0.93	47.19 ± 1.53	59.03 ± 0.34	68.86 ± 0.56	30.93 ± 1.43	41.40 ± 2.46
Qwen-STEM	**41.63 ± 2.10**	52.50 ± 2.14	64.71 ± 1.05	69.69 ± 0.73	31.75 ± 2.81	43.79 ± 1.71
Qwen-Math	39.47 ± 1.66	53.53 ± 1.24	**66.93 ± 0.72**	**74.00 ± 0.59**	30.18 ± 1.65	41.17 ± 2.32
Qwen-BOTH	40.81 ± 2.04	**54.46 ± 1.27**	65.71 ± 0.74	71.64 ± 1.16	**33.83 ± 2.59**	**43.90 ± 2.71**
Llama	29.06 ± 1.44	43.57 ± 0.88	47.73 ± 0.89	60.53 ± 1.67	34.40 ± 2.58	42.27 ± 1.60
Llama-STEM	38.95 ± 1.31	53.17 ± 1.15	59.14 ± 0.85	68.19 ± 1.15	36.04 ± 3.98	46.87 ± 1.49
Llama-Math	36.16 ± 2.33	53.75 ± 1.15	59.65 ± 0.98	69.01 ± 0.55	34.78 ± 4.26	45.55 ± 0.68
Llama-BOTH	**39.43 ± 2.00**	**54.73 ± 1.75**	**63.81 ± 0.90**	**72.74 ± 0.26**	**36.61 ± 2.73**	**48.65 ± 0.49**

Problem Statement. Observing considerable improvements from adding external knowledge ingredients from DeepSeek-R1 to base models in RQ1, we hypothesize similar improvements would scale on reasoning-enhanced models, offering additional gains on top of enhanced reasoning capabilities. To this end, we evaluate base and CoT SFTed variants on KIs extracted from DeepSeek-R1, providing the same necessary knowledge extracted from DeepSeek-R1's reasoning traces (w/ R1 KIs). As a baseline without the added knowledge, we provide the tested models with KIs extracted from their own CoT traces (w/ self KIs) for comparison.

Answer to RQ2: Yes. the reasoning models also substantially benefit from addition of contextual knowledge. As shown in Table 3, within both Qwen and Llama groups, reasoning-enhanced models w/ R1 KIs in the context show significant improvements over the base setting without the KIs, while preserving the gap compared with the base model w/ R1 KIs. Confirming the effectiveness of providing external knowledge as an in-context prompt, this result sheds light on potential future improvement by applying high-quality external memory modules as an external knowledge source for better problem-solving capabilities, echoing the finding in COMPACTDB [28], a concurrent effort constructing a high-quality datastore for reasoning-intensive tasks.

Note, however, that in these experiments, we do not distinguish between two possible non-exclusive explanations for the improvement from adding R1 KIs. (a) It may be that the R1 KIs provide new key

knowledge absent from the model's parameters, or (b) the model may already possess these facts but struggle to retrieve them (put another way, once a strong reasoning model supplies the *key* facts, the reasoning search space might narrow and the problem becomes easier, whether or not the model originally "knew" the augmented facts). We further analyze this confounder in RQ3.

4.2.4 RQ3: Does reasoning fine-tuning improve models' ability to surface helpful knowledge?

Problem Statement. While we observe that external knowledge benefits reasoning models, in this RQ, we ask how reasoning-fine-tuning affects knowledge recall. To this end, we focus on evaluating the KIs from -Math models to determine whether they offer more improvement than those of base models, since -Math models are fine-tuned on math-only data without additional scientific knowledge.

Notably, in Table 2, while -STEM and -BOTH variants, trained with SYNTHETIC-1-STEM, outperform -Math variants due to science in-domain training data, -Math variants also largely outperform the base model even without being trained on science data. Recalling our discussion in RQ2 (§4.2.3), the -Math model's gains have the same two non-exclusive explanations, (a) some science questions require mathematical knowledge, and the -Math model performs better on these because math knowledge was loaded into the model through the math-specific fine-tuning, and/or (b) the -Math model is better at surfacing its relevant parametric knowledge via CoT expression.

Table 4: Accuracy (%) of synthetic knowledge recall on KIs generated from Qwen/Llama-Math on GPQA and MMLU-Pro*. Base models and math reasoning-fine-tuned models show similar performance on knowledge recall questions.

KI Dataset	**Qwen** *Qwen-Math*	**-Math**	**Llama** *Llama-Math*	**-Math**
KI-GPQA	72.30	73.02	70.94	68.94
KI-MMLU-Pro*	82.49	81.50	74.46	74.12

To disentangle these two factors, we extract KIs from the CoTs of the -Math models and examine whether these KIs represent new knowledge added by fine-tuning, or whether they are also facts known to the base model. We probe this by querying the model with synthetic questions that test knowledge of each KI (see Appendix F.2 for examples). Then, to verify explanation (b), we provide the KIs in-context from either the -Math or base model, to the corresponding base model; i.e., holding mathematical reasoning capacity constant while varying only the external knowledge.

Answer to RQ3: Yes. In response to explanation (a), we find that on average, the base models and their corresponding -Math variants have similar recall of the KIs (Table 4), meaning that explanation (a) is unlikely to be the major contributor for the improvements.

To verify explanation (b), Table 5 shows that KIs from -Math deliver significant boosts over those from the base models across different benchmarks and model families. This result suggests that CoT verbalization improves the model's ability to surface the most relevant knowledge for the given reasoning problems. Notably, the KIs are unlikely to have been newly acquired during fine-tuning (Table 4); instead, the findings indicate that reasoning-fine-tuned models exhibit improved recall of knowledge already parameterized in the base model.

Table 5: Performance on GPQA and MMLU-Pro* with KIs extracted from base and -Math reasoning models. KIs extracted from -Math models enable more improvement over those from base models.

Base	**Setup**	**GPQA**	**MMLU-Pro***
Qwen	w/ Qwen KIs	34.24 ± 0.93	59.03 ± 0.34
	w/ Qwen-Math KIs	**36.93 ± 1.75**	**63.66 ± 0.45**
Llama	w/ Llama KIs	29.06 ± 1.44	47.73 ± 0.89
	w/ Llama-Math KIs	**29.69 ± 1.72**	**53.91 ± 0.94**

5 Conclusion

In this work, we studied how reasoning and domain knowledge each contribute to scientific reasoning in LLMs. To this end, we introduced SCIREAS and SCIREAS-PRO, a unified, reproducible suite for evaluating scientific reasoning across domains and formats, together with KRUX, a knowledge-controlled evaluation framework that assesses LLMs using provided knowledge ingredients (KIs). We showed: (i) retrieving task-relevant knowledge from parameters is a key bottleneck; (ii) reasoning-fine-tuned models get complementary gains from external KIs; and (iii) verbalized CoT improves knowledge surfacing. Our results show that reasoning-focused fine-tuning improves both reasoning and knowledge use, suggesting promising future directions in better understanding and enhancing these interconnected components.

References

[1] I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, and A. Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025. URL https://arxiv.org/abs/2503.08679.

[2] A. M. Bean, S. Hellsten, H. Mayne, J. Magomere, E. A. Chi, R. A. Chi, S. A. Hale, and H. R. Kirk. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=cLga8GStdk.

[3] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3620. Association for Computational Linguistics, 2019.

[4] A. Bercovich, I. Levy, I. Golan, M. Dabbah, R. El-Yaniv, O. Puny, I. Galil, Z. Moshe, T. Ronen, N. Nabwani, I. Shahaf, O. Tropp, E. Karpas, R. Zilberstein, J. Zeng, S. Singhal, A. Bukharin, Y. Zhang, T. Konuk, G. Shen, A. S. Mahabaleshwarkar, B. Kartal, Y. Suhara, O. Delalleau, Z. Chen, Z. Wang, D. Mosallanezhad, A. Renduchintala, H. Qian, D. Rekesh, F. Jia, S. Majumdar, V. Noroozi, W. U. Ahmad, S. Narenthiran, A. Ficek, M. Samadi, J. Huang, S. Jain, I. Gitman, I. Moshkov, W. Du, S. Toshniwal, G. Armstrong, B. Kisacanin, M. Novikov, D. Gitman, E. Bakhturina, J. P. Scowcroft, J. Kamalu, D. Su, K. Kong, M. Kliegl, R. Karimi, Y. Lin, S. Satheesh, J. Parmar, P. Gundecha, B. Norick, J. Jennings, S. Prabhumoye, S. N. Akter, M. Patwary, A. Khattar, D. Narayanan, R. Waleffe, J. Zhang, B.-Y. Su, G. Huang, T. Kong, P. Chadha, S. Jain, C. Harvey, E. Segal, J. Huang, S. Kashirsky, R. McQueen, I. Putterman, G. Lam, A. Venkatesan, S. Wu, V. Nguyen, M. Kilaru, A. Wang, A. Warno, A. Somasamudramath, S. Bhaskar, M. Dong, N. Assaf, S. Mor, O. U. Argov, S. Junkin, O. Romanenko, P. Larroy, M. Katariya, M. Rovinelli, V. Balas, N. Edelman, A. Bhiwandiwalla, M. Subramaniam, S. Ithape, K. Ramamoorthy, Y. Wu, S. V. Velury, O. Almog, J. Daw, D. Fridman, E. Galinkin, M. Evans, S. Ghosh, K. Luna, L. Derczynski, N. Pope, E. Long, S. Schneider, G. Siman, T. Grzegorzek, P. Ribalta, M. Katariya, C. Alexiuk, J. Conway, T. Saar, A. Guan, K. Pawelec, S. Prayaga, O. Kuchaiev, B. Ginsburg, O. Olabiyi, K. Briski, J. Cohen, B. Catanzaro, J. Alben, Y. Geifman, and E. Chung. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

[5] A. Bercovich, I. Levy, I. Golan, M. Dabbah, R. El-Yaniv, O. Puny, I. Galil, Z. Moshe, T. Ronen, N. Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

[6] R. Cohen, M. Geva, J. Berant, and A. Globerson. Crawling the internal knowledge-base of language models, 2023. URL https://arxiv.org/abs/2301.12810.

[7] DeepSeek-AI. Deepseek-r1: Usage recommendations, 2025. URL https://huggingface.co/deepseek-ai/DeepSeek-R1#usage-recommendations.

[8] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li,

Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[9] H. Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

[10] K. Feng, K. Ding, W. Wang, X. Zhuang, Z. Wang, M. Qin, Y. Zhao, J. Yao, Q. Zhang, and H. Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2024. URL https://arxiv.org/abs/2406.09098.

[11] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

[12] D. Gottesman and M. Geva. Estimating knowledge in large language models without generating a single token, 2024. URL https://arxiv.org/abs/2406.12673.

[13] J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk, A. Vahdat, P. Kohli, Y. Matias, A. Carroll, K. Kulkarni, N. Tomasev, Y. Guan, V. Dhillon, E. D. Vaishnav, B. Lee, T. R. D. Costa, J. R. Penadés, G. Peltz, Y. Xu, A. Pawlosky, A. Karthikesalingam, and V. Natarajan. Towards an ai co-scientist, 2025. URL https://arxiv.org/abs/2502.18864.

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Y. Gu, O. Tafjord, B. Kuehl, D. Haddad, J. Dodge, and H. Hajishirzi. Olmes: A standard for language model evaluations, 2025. URL https://arxiv.org/abs/2406.08446.

[16] C. He, R. Luo, Y. Bai, S. Hu, Z. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

[17] M. Jin, W. Luo, S. Cheng, X. Wang, W. Hua, R. Tang, W. Y. Wang, and Y. Zhang. Disentangling memory and reasoning ability in large language models, 2025. URL https://arxiv.org/abs/2411.13504.

[18] J. Khouja, K. Korgul, S. Hellsten, L. Yang, V. Neacsu, H. Mayne, R. Kearns, A. Bean, and A. Mahdi. Lingoly-too: Disentangling reasoning from knowledge with templatised orthographic obfuscation, 2025. URL https://arxiv.org/abs/2503.02972.

[19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

[20] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[21] N. Lambert, L. Tunstall, N. Rajani, and T. Thrush. Huggingface h4 stack exchange preference dataset, 2023. URL https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences.

[22] J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnapati, A. D. White, and S. G. Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

[23] J. Lee, W. Yoon, S. Kim, D. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[24] A. O. Li and T. Goyal. Memorization vs. reasoning: Updating llms with new knowledge, 2025. URL https://arxiv.org/abs/2504.12523.

[25] S. Li, J. Huang, J. Zhuang, Y. Shi, X. Cai, M. Xu, X. Wang, L. Zhang, G. Ke, and H. Cai. Scilitllm: How to adapt llms for scientific literature understanding, 2025. URL https://arxiv.org/abs/2408.15545.

[26] Q. Liu, S. Zhang, G. Qin, T. Ossowski, Y. Gu, Y. Jin, S. Kiblawi, S. Preston, M. Wei, P. Vozila, T. Naumann, and H. Poon. X-reasoner: Towards generalizable reasoning across modalities and domains, 2025. URL https://arxiv.org/abs/2505.03981.

[27] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

[28] X. Lyu, M. Duan, R. Shao, P. W. Koh, and S. Min. Frustratingly simple retrieval improves challenging, reasoning-intensive benchmarks, 2025. URL https://arxiv.org/abs/2507.01297.

[29] R. Ma, P. Wang, C. Liu, X. Liu, J. Chen, B. Zhang, X. Zhou, N. Du, and J. Li. S^2r: Teaching llms to self-verify and self-correct via reinforcement learning, 2025. URL https://arxiv.org/abs/2502.12853.

[30] X. Ma, Q. Liu, D. Jiang, G. Zhang, Z. Ma, and W. Chen. General-reasoner: Advancing llm reasoning across all domains. 2025. URL https://api.semanticscholar.org/CorpusID:278768680.

[31] J. Mattern, F. Gabriel, and J. Hagemann. Synthetic-1 release: Two million collaboratively generated reasoning traces from deepseek-r1, February 2025. URL https://www.primeintellect.ai/blog/synthetic-1-release.

[32] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

[33] OpenAI, :, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer,

M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, and Z. Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

[34] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models, 2023. URL https://arxiv.org/abs/2309.00071.

[35] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel. Kilt: a benchmark for knowledge intensive language tasks, 2021. URL https://arxiv.org/abs/2009.02252.

[36] A. Prabhakar et al. Omniscience: A domain-specialized llm for scientific reasoning. *arXiv preprint arXiv:2503.17604*, 2025.

[37] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

[38] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, Z. Liu, and E. Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

[39] L. Sun, Y. Han, Z. Zhao, D. Ma, Z. Shen, B. Chen, L. Chen, and K. Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research, 2024. URL https://arxiv.org/abs/2308.13149.

[40] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[41] P. Team, X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, C. Zheng, K. Deng, S. Gavin, S. Jia, S. Jiang, Y. Liao, R. Li, Q. Li, S. Li, Y. Li, Y. Li, D. Ma, Y. Ni, H. Que, Q. Wang, Z. Wen, S. Wu, T. Hsing, M. Xu, Z. Yang, Z. M. Wang, J. Zhou, Y. Bai, X. Bu, C. Cai, L. Chen, Y. Chen, C. Cheng, T. Cheng, K. Ding, S. Huang, Y. Huang, Y. Li, Y. Li, Z. Li, T. Liang, C. Lin, H. Lin, Y. Ma, T. Pang, Z. Peng, Z. Peng, Q. Qi, S. Qiu, X. Qu, S. Quan, Y. Tan, Z. Wang, C. Wang, H. Wang, Y. Wang, Y. Wang, J. Xu, K. Yang, R. Yuan, Y. Yue, T. Zhan, C. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhao, X. Zheng, C. Zhong, Y. Gao, Z. Li, D. Liu, Q. Liu, T. Liu, S. Ni, J. Peng, Y. Qin, W. Su, G. Wang, S. Wang, J. Yang, M. Yang, M. Cao, X. Yue, Z. Zhang, W. Zhou, J. Liu, Q. Lin, W. Huang, and G. Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

[42] Q. Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

[43] Q. Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

[44] R. Thapa, Q. Wu, K. Wu, H. Zhang, A. Zhang, E. Wu, H. Ye, S. Bedi, N. Aresh, J. Boen, S. Reddy, B. Athiwaratkun, S. L. Song, and J. Zou. Disentangling reasoning and knowledge in medical large language models. *ArXiv*, abs/2505.11462, 2025. URL https://api.semanticscholar.org/CorpusID:278714970.

[45] A. Turpin, J. Wei, D. Zhou, Q. V. Le, and E. H. Chi. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2305.15020*, 2023.

[46] D. Wadden, K. Shi, J. Morrison, A. Naik, S. Singh, N. Barzilay, K. Lo, T. Hope, L. Soldaini, S. Z. Shen, D. Downey, H. Hajishirzi, and A. Cohan. Sciriff: A resource to enhance language model instruction-following over scientific literature, 2024. URL https://arxiv.org/abs/2406.07835.

[47] D. Wadden, K. Shi, J. Morrison, A. Naik, S. Singh, N. Barzilay, K. Lo, T. Hope, L. Soldaini, S. Z. Shen, et al. Sciriff: A resource to enhance language model instruction-following over scientific literature. *arXiv preprint arXiv:2406.07835*, 2024.

[48] C. Wang, W. Su, Q. Ai, Y. Zhou, and Y. Liu. Decoupling reasoning and knowledge injection for in-context knowledge editing, 2025. URL https://arxiv.org/abs/2506.00536.

[49] P. Wang et al. Scienceqa: A large-scale open dataset for question answering in science education. *arXiv preprint arXiv:2210.08127*, 2023.

[50] W. Wang, X. Chen, et al. Evaluating the faithfulness of chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2401.02392*, 2024.

[51] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, and W. Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[52] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

[53] Y. Wang, D. Yu, Q. Zhou, et al. Can large language models follow chain-of-thought prompts faithfully? In *International Conference on Learning Representations (ICLR)*, 2024.

[54] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[55] J. Wu, S. Liu, H. Tu, H. Yu, X. Huang, J. Zou, C. Xie, and Y. Zhou. Knowledge or reasoning? a close look at how llms think across domains, 2025. URL https://arxiv.org/abs/2506.02126.

[56] X. Xu, Q. Xu, T. Xiao, T. Chen, Y. Yan, J. Zhang, S. Diao, C. Yang, and Y. Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models, 2025. URL https://arxiv.org/abs/2502.00334.

[57] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[58] X.-W. Yang, X.-Y. Zhu, W.-D. Wei, D.-C. Zhang, J.-J. Shao, Z. Zhou, L.-Z. Guo, and Y.-F. Li. Step back to leap forward: Self-backtracking for boosting reasoning of language models, 2025.

[59] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[60] G. Zhang et al. Sciglm: Pre-training generalist language models for science with scientific papers. *arXiv preprint arXiv:2402.00730*, 2024.

[61] W. X. Zhao et al. A survey of llms for scientific research. *arXiv preprint arXiv:2307.07927*, 2023.

[62] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A Limitations

Our KRUX framework and KI extraction methods depend on strong models like DeepSeek-R1 for generating reasoning traces. While we used an open-weight model, which provides more transparency and interpretability, the KI extraction pipeline may introduce unobservable biases (though risk is minimal due to our focus on scientific domains), unwanted leakage of information about the answer, or inconsistencies in the faithfulness of the KIs to the task. To mitigate this, we conducted manual analysis of the KIs, confirming their relevance and no direct answer leakage, but extracted KIs could occasionally be irrelevant or incomplete, especially if deployed at scale. Furthermore, some of our analyses are confounded by factors such as context sensitivity (addressed via permutations) and the impact of constraining the search space when providing KIs, which we interpret as an upper bound but may overestimate pure recall benefits. We have taken measures to mitigate these and discussed the caveats in our discussion of results with more details.

Our experiments focus on moderate-sized LLMs with <10B parameters, specifically open-weight models (Qwen2.5, Llama3.1). While we deliberately selected two model families and models large enough to exhibit non-trivial reasoning performance, this limits the generalizability of our findings to larger models. Experimenting with larger models represents a straightforward extension but requires significantly greater computational resources, beyond the scope of our current study and our available compute resources.

The benchmarks we examine emphasize STEM fields, which may underrepresent interdisciplinary or emerging scientific research areas. We acknowledge potential data contamination issues that may impact our analysis; however, the nature of our study is analytical, and we perform controlled experiments. In our benchmarks, we also mitigate these concerns by focusing on recent 2024–2025 datasets. Despite these constraints, our methodology provides a systematic framework for evaluating domain-specific reasoning that can be extended to address these limitations in future work.

B Extended Related Work

Evaluating Knowledge of LLMs Early efforts tended to evaluate the LM knowledge frontier with a static unified benchmark [35]. However, given the growing training corpus for pushing LLM performance, quantifying the knowledge frontier of LLMs becomes increasingly challenging, making it difficult to design a unified benchmark. Instead of general knowledge evaluation, recent work approaches the knowledge frontier of LLMs by anchoring on specific entities, proposing methods to quantify knowledge and factuality around given entities [12, 6]. With recent development of reasoning LLMs, more work exploits long CoT traces as evidence of explicit knowledge utilization, verifying knowledge recall in CoT traces for factuality [55]. Nevertheless, directly evaluating CoT traces can result in false positive signals on the knowledge boundary, given that the knowledge involved could be factual but not helpful for problem solving [1]. In our framework, we construct controlled settings and protocols to evaluate whether the knowledge is genuinely helpful for problem-solving, implicitly guaranteeing the factuality and relevance.

Reasoning LLMs Recent work has shown that LLMs can be trained to utilize intermediate tokens for reasoning, achieving better performance on reasoning tasks as the decoding budget increases. OpenAI's o-series [33] represents the landmark of this paradigm among commercial frontier models, followed by DeepSeek-R1 [8] and several recent efforts to reproduce this success without releasing the training data, such as QwQ [43] and Kimi [40]. Some recent initiatives aim to achieve the same goal using fully open data sources, led by Llama-Nemotron from NVIDIA [5] and SYNTHETIC-1 from Prime Intellect [31], releasing post-training data to foster development within the community. Our work builds on these commitments, sharing the vision of improving model reasoning by leveraging intermediate tokens, while emphasizing our focus on scientific domains rather than on mathematics or general logical reasoning.

LLMs for Science Recent advancements in scientific LLMs have transitioned from early domain-specific pretraining (e.g., Beltagy et al. 3, Lee et al. 23), to more comprehensive models with multiple stages of training, e.g., SciGLM [60], SciLitLLM [25], and OmniScience [36]. On the other hand, reasoning models have shown strong performance on scientific tasks such as GPQA and MMLU-Pro [8, 33], and some recent efforts instrument LLMs to separate recall from deduction during inference

[48, 17]. However, we still lack a clear understanding of the factors underlying performance on scientific tasks, such as knowledge acquisition or improved reasoning capabilities. We aim to address this gap by studying these factors and then providing a recipe for training more capable models in science.

C SCIREAS Details

Benchmark	o3			o3-mini			o4-mini			Gemini-2.5-Pro			Claude-Sonnet-4			GPT-5		
	Low	High	Δ	Low	High	Δ	Low	High	Δ	Low	High	Δ	Low	High	Δ	Low	High	Δ
GPQA	75.4	79.9	+4.5	63.4	73.9	+10.5	69.4	74.6	+5.2	80.1	79.5	-0.6	63.8	69.0	+5.2	79.2	82.4	+3.1
SuperGPQA*	54.9	59.5	+4.6	40.5	54.0	+13.5	48.6	57.1	+8.5	60.1	60.4	+0.3	45.2	49.8	+4.6	58.6	62.4	+3.8
MMLU-Pro*	85.7	86.6	+0.9	82.1	85.0	+2.9	84.1	86.0	+1.9	85.0	86.2	+1.2	84.1	85.3	+1.2	86.5	88.6	+2.1
LabBench*	70.5	74.2	+3.7	56.9	59.2	+2.3	59.7	63.7	+4.0	61.9	64.4	+2.5	53.4	57.2	+3.8	66.6	74.4	+7.8
OlympBench	53.5	58.0	+4.5	39.5	51.1	+11.6	40.4	49.6	+9.2	67.5	69.6	+2.1	55.4	59.8	+4.4	60.0	64.9	+4.8
SciBench	69.7	72.1	+2.4	46.0	66.3	+20.3	65.5	69.7	+4.2	71.0	70.2	-0.8	65.5	67.1	+1.6	70.4	72.0	+1.6
SciEval*	84.8	82.7	-2.1	83.8	83.4	-0.4	87.1	87.5	+0.4	86.4	85.1	-1.3	85.8	85.8	0.0	87.4	86.1	-1.3
SciKnowEval*	52.1	51.9	-0.2	49.0	51.9	+2.9	49.9	51.1	+1.2	46.8	47.6	+0.8	43.6	43.3	-0.3	45.5	46.7	+1.2
SciRIFF*	51.8	53.6	+1.8	51.3	51.8	+0.5	50.6	52.2	+1.6	51.6	51.4	-0.2	53.5	50.9	-2.6	46.9	50.1	+3.3
UGPhysics*	63.1	65.2	+2.1	56.7	60.7	+4.0	57.7	62.2	+4.5	56.0	55.4	-0.6	52.4	53.2	+0.8	63.6	67.6	+4.0
Average	66.2	68.4	+2.2	56.9	63.7	+6.8	61.3	65.4	+4.1	66.6	67.0	+0.4	60.3	62.1	+1.8	66.5	69.5	+3.1
0.01$ / Instance	0.68	2.25	×3.3	0.41	3.24	×7.9	0.41	2.38	×5.8	1.07	12.51	×11.7	1.83	7.50	×4.1	0.72	3.10	×4.3

Table 6: Performance (%) across SCIREAS grouped by models at low and high reasoning efforts. The same model with different reasoning effort can have distinctive performance with a clear margin.

C.1 Evaluation Suite Curation

See Table 11-12 for domain distribution. We list the selection of each benchmark as follows.

GPQA [37]: No change. Report in micro average. License: CC-BY-4.0.

MMLU-Pro [52]: MMLU-Pro features subjects beyond STEM and scientific subjects. We first filter by subjects, retaining instances from physics, chemistry, computer science, math, biology, and health, and then randomly sample each task to 200 instances max. Report in macro average across 7 subjects. License: MIT.

LabBench [22]: We drop tasks that require visual inputs or external table/paper extraction, therefore dropping DbQA, FigQA, LitQA2, SuppQA, and TableQA, retaining CloningScenarios, PropotolQA, and SeqQA. Report in macro average across 3 tasks. License: CC-BY-SA-4.0.

SciBench [51]: No change. Report in micro average. License: MIT.

OlympiadBench [16]: Dropping tasks that require visual inputs or not in English. Report the macro average across math and physics. License: apache-2.0.

SciRIFF [47]: We drop tasks that primarily focus on information/relation/table extraction and retain EvidenceInference, Qasper, and SciFact. Report in macro average of 5 metrics (detailed in Table 11-12) across 3 tasks. License: ODC-BY.

SciKnowEval [10]: The authors introduce scientific tasks in 5 progressive levels from knowledge memorization to application. After manual inspection, we only preserve tasks from the highest level of knowledge application (L5), and cap instances from each task to be 200. Report the macro average across 8 tasks. License: MIT.

SciEval [39]: Similar to SciKnowEval, the authors introduce 4 progressive levels of static tasks, including basic knowledge, knowledge application, scientific calculation, and research creativity. After inspection, we retain knowledge application and scientific calculation subsets, capping each task to a maximum of 200. Report the macro average across 6 tasks. License: N/A.

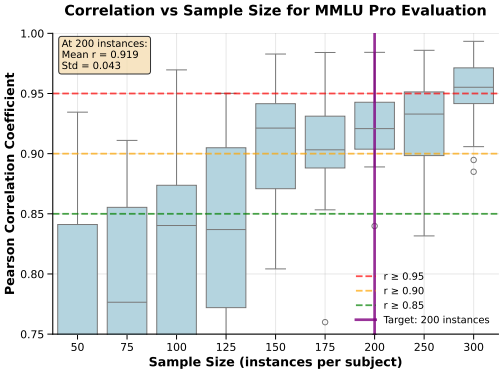

Figure 6: Correlation between sampled and full dataset performance as a function of sample size. 200 instances per subject (purple) yields r = 0.919 ± 0.043. Error bars: SD over 30 samples.

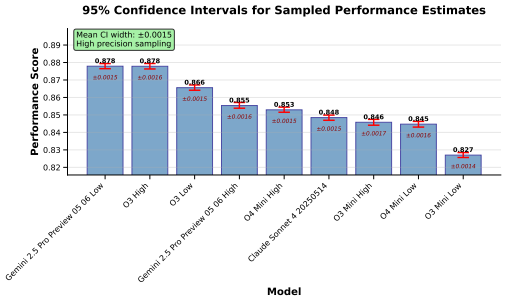

Figure 7: 95% confidence intervals for 200-instance sampling across nine frontier models. Mean CI half-width ≈ 0.0015; numbers above/below bars show mean and half-width.

UGPhysics [56]: The authors annotate each instance into 5 different physics reasoning skills: knowledge recall, laws application, math derivation, practical application, and others, which fails to be categorized into the categories above. We filter out instances that specifically require knowledge recall only and cap instances from each subject to be 200 max. Report the macro average across 13 subjects. License: CC-BY-NC-SA-4.0.

SuperGPQA [41]: We curate questions from two broad domains — science and engineering — while omitting niche areas that lie outside mainstream STEM (e.g., weapon science, textile engineering). The science portion spans mathematics, biology, physics, systems science, and chemistry. The engineering portion covers a comprehensive set of disciplines: electronic science and technology; nuclear science and technology; mechanical engineering; information and communication engineering; civil engineering; instrument science and technology; computer science and technology; control science and engineering; chemical engineering and technology; mechanics; electrical engineering; materials science and engineering; hydraulic engineering; power engineering and engineering thermophysics; and optical engineering. Report in macro average across the domain of science and engineering. License: ODC-BY.

C.2 Uniform Sampling Validation: MMLU-Pro Case Study

Evaluating state-of-the-art frontier models could be expensive. To mitigate evaluation cost, we evaluate frontier models on MMLU-Pro* before and after uniform sampling. By sample size correlation in Figure 6 and 95% confidence intervals for sampled subset in Figure 7, we show that the sampling is cost-efficient and statistically effective while reducing evaluating instances from 6,696 to 1,400.

For costly frontier reasoning models such as Gemini-2.5-Pro-Preview, at rates in time of writing, the sampling reduces SCIREAS evaluation costs from $3,600 to $1,500 and can be further decreased to $730 by using batch job inference.

C.3 SCIREAS-PRO Reasoning Intensiveness Validation

To test this hypothesis, we pursue two complementary checks: (1) different reasoning models should have high agreement identifying reasoning intensive instances, and (2) filtered instances should agree with human judgment in terms of reasoning intensiveness.

C.3.1 Cross-Model Agreement on Reasoning Intensity

To validate our hypothesis that performance gaps between different reasoning effort settings indicate reasoning intensity, we first examine whether different models agree on which instances are reasoning-

intensive. As shown in Figure 1, for each reasoning model, we categorize each test question from SCIREAS by their correctness under low/high reasoning efforts into four categories, `(high_c, low_c)`, `(high_c, low_i)`, `(high_i, low_c)`, and `(high_i, low_i)`, where `high`/`low` stands for high/low reasoning effort setting and `*_c`/`*_i` stands for the problem instance has been answered correctly/incorrectly by the model. Treating `(high_c, low_i)` as targeting instances that require high reasoning effort, we measure how `(high_c, low_i)` sets derived from different reasoning models agree with others.

As shown in Table 7, treating `(high_c, low_i)` from o3-mini as ground truth, the same set derived from o4-mini, o3, and claude-sonnet-4 largely coincide with o3-mini across different benchmarks from SCIREAS (all above 70%), showing high agreement on instances that require high reasoning efforts across models from different model families.

Ground Truth **Target**	**o3-mini vs. o4-mini**	**o3-mini vs. o3**	**o3-mini vs. claude-sonnet-4**
SuperGPQA*	78.0	77.8	76.1
GPQA	80.4	81.0	79.0
MMLU-Pro*	92.2	91.6	92.0
LabBench*	71.9	74.6	75.8
SciBench	75.9	74.1	75.4
OlympiadBench	81.1	81.5	81.2
SciEval*	94.3	93.1	93.5
UGPhysics*	83.2	82.9	83.8

Table 7: Accuracy of overlapping instances on `(high_c, low_i)` from o3-mini vs. other models, treating o3-mini as ground true label. Different reasoning models agree on high reasoning instances.

C.3.2 Human and LLM-as-Judge Assessment

The overlap of instances that require high reasoning effort shows reasoning models tend to agree on problem difficulty, but to verify the reliability of reasoning effort as a surrogate, the filter should also align with human judgment.

To this end, we collect the union of `(high_c, low_i)` from o3-mini and o4-mini for the case study and apply an LLM-as-judge assessment [62] to expedite the process while manually annotating a subset for a reliability test. The LLM judge is based on GPT-4.1 for a balanced tradeoff between assessment reliability and cost. Notably, naively prompting the LLM judge to determine the reasoning difficulty could be suboptimal due to a lack of reference. Therefore, we designed two reference-based evaluation protocols: (a) pair-wise comparison on reasoning difficulty between instance questions sampled from filtered subset and original SCIREAS, and (b) identifying failing reason for filtered instances given low and high reasoning outputs (i.e., whether the model fails in a low reasoning setting due to lack of reasoning effort).

(a) Pairwise Comparison For each instance in SCIREAS-PRO, the judge is also presented with an instance drawn from the set of other, non-overlapping instances from SCIREAS. The judge is not given any information as to which instance is drawn from which source and is tasked to identify which instance is more reasoning-intensive.

(b) Failure Analysis For each instance in SCIREAS-PRO, the judge is presented with both the correct high reasoning output (if both o3-mini-high and o4-mini-high are correct, o4-mini-high will be selected) as well as the incorrect low reasoning output from the corresponding model (e.g. correct: o3-mini-high; incorrect: o3-mini-low). The judge is tasked with determining whether the failure of the low reasoning effort model can be attributed primarily due to insufficient reasoning ability or lack of domain knowledge.

SYSTEM MESSAGE

You are an expert judge comparing reasoning intensity between two questions. Analyze both questions thoroughly and determine which one demands more complex reasoning.
Reply in this exact format:
```
###EXPLANATION: <detailed analysis of both questions and the comparison>
###RESULTS: A / B / UNCLEAR
```

USER MESSAGE

You will be shown two questions (A and B) from the same academic domain.
A question is *reasoning intensive* if it requires:
- Complex multi-step logical reasoning
- Advanced mathematical computation or derivation
- Integration of multiple concepts or principles
- Abstract thinking or sophisticated problem-solving strategies
- Deep domain knowledge application

QUESTION A
Context: `{{context_a}}`
Question: `{{question_a}}`
QUESTION B
Context: `{{context_b}}`
Question: `{{question_b}}`
Analyze both questions carefully and explain your reasoning. Then reply using the exact format specified above.

Figure 8: Full reasoning intensiveness pairwise comparison prompt template used in our experiments.

SYSTEM MESSAGE

You are an expert judge analyzing why AI models fail on reasoning-intensive questions. Compare the correct and incorrect answers to determine if the failure was primarily due to insufficient reasoning ability or lack of domain knowledge.
Reply in this exact format:
```
###EXPLANATION: <detailed analysis of why the low-reasoning model failed>
###RESULTS: REASONING/KNOWLEDGE/BOTH/UNCLEAR
```

USER MESSAGE

You will be shown a question from an academic dataset, along with
(1) a *CORRECT* answer from a high-reasoning model and
(2) an *INCORRECT* answer from a low-reasoning model.
Your task is to analyze *why* the low-reasoning model failed.
Consider whether the failure is primarily due to:
- *REASONING*: Insufficient logical thinking, problem-solving ability, or step-by-step analysis
- *KNOWLEDGE*: Lack of domain knowledge (missing facts, formulas, concepts, procedures)
- *BOTH*: Significant deficiencies in both reasoning and knowledge
- *UNCLEAR*: Cannot determine the primary cause of failure

QUESTION
Context: `{{context}}`
Question: `{{question}}`
CORRECT ANSWER (from `{{high_model}}`):
`{{high_full_response}}`
INCORRECT ANSWER
(from `{{low_model}}`):
`{{low_full_response}}`
Analyze why the low-reasoning model failed. Was it primarily due to insufficient reasoning ability or lack of knowledge?

Figure 9: Prompt used to classify failure cause (reasoning vs. knowledge) for low-reasoning models.

Results We show that both protocols agree that filtered instances require significantly more reasoning efforts than non-filtered instances from SCIREAS, with (a) showing 71% agreement in accuracy by LLMs with 78% human annotation agreement and (b) showing 91% agreement by LLMs with 90% human agreement, where human annotations are made by authors on 80 sampled tests for each protocol.

D Frontier Model API Evaluation Configuration

For OpenAI and xAI provided reasoning models, we apply generic "low" and "high" reasoning effort parameters with respect to official documentation where specificity on token budget is not allowed; for other reasoning models that allows thinking budgets as input (e.g. Gemini and Anthropic), we adopt "low" as definition introduced by LiteLLM,[8] which corresponds to 1024 budget, and remove the constraint to allow for as many thinking tokens as the model needed to unleash full potential as "high" reasoning effort, corresponding to the highest reasoning effort from OpenAI and xAI models. For all frontier reasoning models, if not restricted, we set temperature=1, borrowed from OpenAI forced setting,[9] and top-p=0.95, borrowed from recommended setting by Anthropic,[10] with max generation length of 64K, as we observe no models tend to output more than 20K tokens. We log API pricing at the time of writing in Table 8.

Model	Input Price ($ per 1M tokens)	Output Price ($ per 1M tokens)
OpenAI models		
GPT-4.1-2025-04-14	2.00	8.00
o3-mini-2025-01-31	1.10	4.40
o3-2025-04-16	2.00	8.00
o4-mini-2025-04-16	1.10	4.40
GPT-5-2025-08-07	1.25	10.00
GPT-oss-120B (Together AI)	0.15	0.60
DeepSeek models		
DeepSeek-V3-0324	0.14	0.28
DeepSeek-R1-0120	0.55	2.19
DeepSeek-R1-0528	0.55	2.19
Alibaba Qwen models (Together AI)		
Qwen3-32B	0.40	1.20
Google models		
Gemini-2.5-Pro-Preview-05-06	1.25	10.00
Meta models (Together AI)		
Llama-4-Maverick-17B-128E-Instruct-FP8	0.27	0.85
Anthropic models		
Claude-Sonnet-4-20250514	3.00	15.00

Table 8: Pricing ($ per 1M tokens) for input and output across different LLM providers at the time of writing, without any discounts.

E Training / Evaluation Details

E.1 Distillation from Reasoning LLMs

To obtain high-performing reasoning models for study, we employ a distillation method that fine-tunes smaller models using Supervised Fine-tuning (SFT) on the CoT trajectories generated by large reasoning models, as it is more effective than reinforcement learning (RL) with the small models alone [8]. Specifically, we consider the standard SFT framework for language models where the objective is to train a model f_θ to approximate a distribution over output sequences y conditioned on input x, based on a dataset $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$. For recent reasoning LLMs such as DeepSeek-R1, the output y consists of two main parts: a reasoning trace r and the actual output a.

[8] https://docs.litellm.ai/docs/providers/anthropic#usage—thinking–reasoning_content

[9] https://community.openai.com/t/o3-mini-unsupported-parameter-temperature/1140846/3

[10] https://docs.anthropic.com/en/docs/build-with-claude/extended-thinking#feature-compatibility

In practice, the reasoning traces are enclosed by keywords `<think>` and `</think>`, indicating the start and the end of the reasoning process. The model is trained with the standard SFT objective: $\mathcal{L}(\theta) = -\Sigma_{(x,y)\in\mathcal{D}}\Sigma_{t=1}^{|y|}\log p_\theta(y_t|y_{<t}, x)$, where y_t is the t-th token and $y_{<t}$ is its prefix.

E.2 Extended Setup

E.2.1 Training Settings

We filter out instances with a token length greater than 4096.[11] The models are trained for 5 epochs with a cosine learning rate scheduler, a maximum learning rate of 1e-5, and 3% warmup steps.

E.2.2 Evaluation setup

The reasoning models could produce excessively long outputs, and may be prone to self-repetition with greedy decoding [7]. In this work, unless otherwise specified, we apply greedy decoding on non-CoT fine-tuned models and top-p=0.95, temperature=0.6 on reasoning models, with a maximum generation length of 64K. From our preliminary studies, we observe that the setup generally reflects the best performance for both settings, and the decoding setup matches the recommended setup from recent efforts in large reasoning models, such as Llama-Nemotron [4]. Notably, for Qwen [57] models and their variants, we apply YaRN context extension [34] as recommended by the official model card [42].

E.3 Math vs. Non-Math

As shown in Table 1, Qwen-STEM and Qwen-Math both exhibit significant improvement over the base model on SCIREAS and SCIREAS-PRO. Qwen-Math slightly outperforms Qwen-STEM on SCIREAS and the gap is amplified on SCIREAS-PRO.

Given limited subject coverage on SYNTHETIC-1-Math dataset, the strong performance of checkpoints fine-tuned on it only seems surprising — Does the improvement come from generalization from math reasoning to a wider domain, or is it because the high-reasoning instances in our datasets are math-intensive? To answer this question, we categorize SCIREAS-PRO into math and non-math instances by heuristics. Specifically, we label instances as math-needed if they contain explicit numeric quantities that typically imply computation. Importantly, numbers that appear solely within unit expressions (e.g., "cm^2") or chemical formulas (e.g., "H_2O" or "NaCl") are not treated as indicators of math-related reasoning.

A question is marked *math-containing* when it includes

1. a signed or unsigned integer/decimal (e.g. `3`, `-2.5`, `60`, `9.81`),
2. **not** embedded inside a word (so digits in `H2O`, `COVID-19`, `IL-2...` are ignored), and
3. optionally followed—without intervening letters—by *any one* of the unit strings listed in Fig. 10.

As shown in Table 9, we find that math computation appears frequently among reasoning-intensive instances, and the improvements on SCIREAS-PRO mostly come from improved math capabilities. For non-math instances, -math variants hardly improve, while -STEM variants and -BOTH variants, trained with STEM subjects data, show noticeable improvements.

E.4 Training Knowledge Enhanced Scientific Reasoning Models

Our post-trained checkpoints are based on models fine-tuned on either SYNTHETIC-1-Math, SYNTHETIC-1-STEM, or both, while combining the two, which cover both STEM and mathematical reasoning, achieves the strongest performance (Table 1). To further assess the effectiveness of this Math+STEM data mixture following §4.1, we compare it directly against concurrently released long-CoT SFT datasets on the same base model. We then apply the same mixture to Qwen3-8B-Base to obtain SCILIT01 to provide a stronger baseline.

Specifically, we compare Qwen-BOTH, which is fine-tuned using our training recipe, with SYNTHETIC-1-SFT [31], a model fine-tuned on SYNTHETIC-1 with additional coding and preference alignment data, and Qwen-Nemotron, a model we trained with the same settings and same amount of

[11] Longer input lengths would slow down our training in quadratic order based on 8 80GB A100/H100 GPUs.

Units recognised by the heuristic

- % °C, °F, K, °
- g, kg, mg, μg/ug, lb/lbs, oz
- m, cm, mm, km L/l, mL/ml, μL/μl/ul
- Pa, kPa, MPa, atm, bar, mbar
- J, kJ, MJ; W, kW, MW, GW
- V, kV; A, mA, μA/uA
- Hz, kHz, MHz, GHz
- cm^2, m^2, mm^2, km^2 cm^3, m^3, mm^3, km^3
- mol; M, mM, μM/uM, nM, pM
- dB; rpm; rad/s
- s, ms, μs/us, ns; min, h day/days; yr/yrs

Figure 10: Unit suffixes accepted by the numeric heuristic. A standalone number with any of these units (or no unit) is treated as evidence that the question contains mathematical content.

Model	Math Acc.	Non-Math Acc.
SCIREAS-PRO*: 1,260 Instances*		
#	1,172	88
Qwen	14.25	12.50
Qwen-STEM	15.53	23.86
Qwen-Math	17.58	13.64
Qwen-BOTH	20.56	28.41
Llama	11.52	13.64
Llama-STEM	14.16	15.91
Llama-Math	17.24	13.64
Llama-BOTH	15.96	23.86

Table 9: Accuracy breakdown on math and non-math instances for SCIREAS-PRO. -Math and -STEM variants contribute to different dimensions of performance, while -BOTH captures improvements on both.

data (§4.1) sampled from science and math domains of Llama-Nemotron [5], a training data mixture for reasoning fine-tuning, all post-trained on Qwen2.5-7B-Instruct. The results in Table 10 show that our data composition yields a stronger baseline for scientific reasoning than concurrent data recipes on Qwen2.5-7B-Instruct (Table 10 center block), and Qwen-BOTH reaches comparable performance to models from concurrent efforts focusing on reasoning enhancement post-training recipes (Table 10 left-hand block, i.e., OpenR1 [9], Llama-Nemotron [5], and General-Reasoner [30]).

Furthermore, using our recipe, we fine-tune the recently released Qwen3-8B-Base to deliver a stronger model, SCILIT01. While its performance falls behind Qwen3-8B with the thinking mode, which has undergone more sophisticated post-training, it outperforms Qwen3-8B with non-thinking mode (Table 10 right-hand block). This indicates that SCILIT01 partially unleashes the reasoning capabilities from the base model, offering a strong baseline for future study on post-training recipe for scientific reasoning.

F Extended KRUX Details

F.1 Knowledge Extraction

In this work, we apply DeepSeek-R1 as the extractor. Prompt shown in Figure 11. We show a set of KIs extracted from Qwen2.5-7B-Instruct (Figure 12), Qwen-Math variants (Figure 13), and DeepSeek-R1 (Figure 14) for the same question from GPQA:

```
Question: A large gene has dozens of exons, of which the central ones
code for folded triple helical repeats that connect the cytoskeleton with
sarcolemma and extracellular space. Each exon usually codes for one folded
triple alpha helix. The most common mutations of the gene are central exon
deletions that create out-of-frame peptides and progressive degenerative
organ waste. A solution is to deliver a Morpholino that recognizes the 5'
end of the out-of-frame exon in pre-mRNA. The molecule prevents binding of
the spliceosome and creates exon skipping and in-frame joining. Several
missing exons are well tolerated by an organism. Which structure below is
not involved in the proposed therapy? (A) lariat (B) antisense (C) R-loops
(D) polyA tail.
```

Models	OpenR1	Llama-Nemotron	General-Reasoner	SYNTHETIC-1-SFT	Qwen-Nemotron	Qwen-BOTH	SCILIT01	Qwen3	Qwen3-thinking
Base Model	Q2.5-Math	L3.1-Inst.	Q2.5-Base	Q2.5-Inst.			Q3-Base		
Training Methods	SFT	SFT&RL	RL	SFT	SFT	SFT	SFT	–	–
Trained by Us	No	No	No	No	Yes	Yes	Yes	No	No
GPQA	44.42	37.95	35.94	38.84	44.20	40.63	50.89	55.80	55.80
SuperGPQA*	31.90	29.39	14.26	22.39	19.47	20.33	30.11	23.32	38.27
MMLU-Pro*	60.86	65.64	62.14	56.21	63.57	65.00	76.57	73.36	81.71
LabBench*	27.14	27.78	35.58	28.61	35.76	33.00	35.07	36.99	38.19
OlympiadBench	53.03	37.62	19.82	40.75	29.33	34.55	43.78	28.51	21.30
SciBench	61.85	57.66	19.08	51.59	48.27	47.11	61.27	54.05	68.21
SciEval*	43.64	68.67	70.34	46.41	38.53	72.36	80.60	81.51	84.02
SciKnowEval*	28.45	30.69	34.19	19.13	31.85	32.00	39.46	37.99	41.81
SciRIFF*	29.17	34.01	37.75	28.57	39.24	41.81	44.01	47.23	47.26
UGPhysics*	50.30	45.92	20.86	43.96	46.52	40.03	52.28	30.98	59.81
Average	43.08	43.53	34.99	37.64	39.67	42.68	<u>51.41</u>	46.97	**53.64**
SCIREAS-PRO	<u>26.43</u>	23.75	13.73	19.44	19.68	21.11	24.84	19.05	**29.92**

Table 10: Performance of concurrent efforts on open-recipe post-training in <10B-parameter level. SCILIT01 shows competitive performance relative to concurrent reasoning post-training methods. We abbreviate Qwen2.5, Qwen3, and Llama-3.1 as Q2.5, Q3, and L3.1, respectively; '-Inst.' denotes the instruction-tuned variant. The best and second-best overall results are highlighted in bold and underlined, respectively.

USER MESSAGE

You are given a reasoning chain that explains and justifies a particular conclusion or answer. Your task is to extract **all distinct knowledge pieces** from this chain. A knowledge piece is any standalone statement conveying an explicit fact, definition, mechanism, relationship, or insight that can be generalized beyond the specific question.

Instructions:

1. Read the entire reasoning chain.
2. Identify each discrete fact or insight expressed.
3. Rewrite each as a self-contained, generalizable sentence.
4. Do **not** include any contextual or example-specific details.
5. Output **only** a list of those sentences.

Output Format:

- knowledge-Piece-1
- knowledge-Piece-2
- ...

Reasoning Chain: `{{REASONING}}`

Now perform the extraction.

Figure 11: Full knowledge extraction prompt template used in our experiments.

KNOWLEDGE EXTRACTION OUTPUT

Based on the reasoning chain, here are all distinct knowledge pieces:
1. R-loops are structures formed when a strand of RNA hybridizes with its complementary DNA strand, displacing the other DNA strand.
2. Morpholino can prevent spliceosome binding.
3. Preventing spliceosome binding can lead to exon skipping.
4. Exon skipping can result in in-frame joining of exons.
5. The process of using Morpholino to induce exon skipping does not involve R-loops.

Figure 12: Example KIs extracted from Qwen2.5-7B-Instruct.

KNOWLEDGE EXTRACTION OUTPUT

Based on the reasoning chain, here are all distinct knowledge pieces:
1. Morpholino oligonucleotides can bind to the 5' splice site of pre-mRNA to block spliceosome recognition.
2. Blocking spliceosome binding prevents normal intron removal and leads to exon skipping.
3. Exon skipping can enable in-frame joining of adjacent exons, preserving the reading frame.
4. Antisense molecules function by complementary base-pairing to specific RNA sequences to modulate splicing.
5. Lariat structures are formed during canonical splicing via branch site interactions but are absent when spliceosome activity is inhibited.
6. R-loops are DNA-RNA hybrid structures unrelated to spliceosome-mediated exon skipping mechanisms.
7. The polyA tail is added post-transcriptionally and is not directly involved in splice site selection or exon skipping processes.

Figure 13: Example KIs extracted from Qwen-Math.

KNOWLEDGE EXTRACTION OUTPUT

Based on the reasoning chain, here are all distinct knowledge pieces:
1. Morpholinos are antisense oligonucleotides that bind to complementary RNA sequences.
2. The spliceosome recognizes splice sites on pre-mRNA through components like U1 snRNP, which binds to the 5' splice site.
3. Blocking spliceosome access to a splice site prevents recognition of exon boundaries and causes exon skipping.
4. Exon skipping can restore the reading frame by joining exons flanking the skipped exon.
5. Functional proteins may be produced if skipped exons encode non-critical regions and the remaining exons maintain an in-frame sequence.
6. Lariat structures form during canonical splicing when introns are removed via branch site cleavage.
7. R-loops are nucleic acid structures involving RNA-DNA hybrids and are not directly involved in spliceosome-mediated splicing.
8. Antisense oligonucleotide therapies rely on sequence-specific binding to pre-mRNA targets.
9. Polyadenylation (polyA tail addition) occurs during mRNA maturation and is unrelated to splice site selection or exon skipping mechanisms.

Figure 14: Example KIs extracted from DeepSeek-R1.

F.2 Knowledge Probing

We provide our probing question synthesis prompt (Figure 15), example input and output (Figure 16), and knowledge probing results in Table 4.

USER MESSAGE

You are a meticulous question-authoring assistant. Your job is to convert declarative knowledge statements into *probing* multiple-choice questions that can test whether another language model truly stores the fact in its parametric memory.
IMPORTANT INSTRUCTIONS FOR QUESTIONS:
1. Factual: It should be about a specific detail or fact mentioned in the statement. For example, a true or false statement, a statistic, a definition, etc.
2. Important: It should be a question about the main topic or a key detail/finding/conclusion of the statement.
3. Context-Independent: It should be fully understandable on its own, without phrases like "the proposed model" or "this approach" that assume prior context.

IMPORTANT INSTRUCTIONS FOR ANSWERS:
1. Provide one correct answer and 4 - 6 incorrect answers.
2. Ensure all answers are roughly the same length and follow a similar style so the correct answer cannot be guessed based on length or style alone.
3. The incorrect answers must be plausible but ultimately wrong, reflecting a misunderstanding or misinterpretation of the knowledge.

OUTPUT FORMAT: Please provide the question, correct answer, incorrect answers, and a list of text snippets from the article as "evidences" in the following format:
{ "question": "Your question here",
"correct_answer": "Correct answer here",
"incorrect_answers": ["Incorrect answer 1", ..., "Incorrect answer N"],
"evidences": ["Text snippets from the article that supports the question and correct answer", "Another text snippet"]
}
# Knowledge Statement: `{src_text}`
Please provide your response in the specified format without any additional text.

Figure 15: Knowledge probing question synthesis template used in our experiments.

EXAMPLE `src_text`

"Hyperfine structure in EPR spectroscopy arises from interactions between unpaired electrons and nuclear spins."

EXAMPLE OUTPUT

{
"question": "What causes hyperfine structure in EPR spectroscopy?",
"correct_answer": "Interactions between unpaired electrons and nuclear spins",
"incorrect_answers": [
"Interactions between electron spins and lattice vibrations", "Coupling between electron orbitals and magnetic fields", "Dipolar interactions between neighboring nuclei", "Spin-orbit coupling within the electron cloud", "Chemical shift anisotropy of atomic orbitals"],
"evidences": [
"Hyperfine structure in EPR spectroscopy arises from interactions between unpaired electrons and nuclear spins."]
}

Figure 16: Knowledge probing question synthesis example input and output.

G LLM Usage Statement

We used ChatGPT-o3 from OpenAI for grammar and typo corrections.

Domain	Task Source	Subtask/Subdomain	Instances	Total	Metrics
Physics	GPQA	Physics	187	5087	Acc
	MMLU-Pro	physics	200		Acc
	SciBench	fund	81		Acc
		thermo	83		Acc
		class	63		Acc
	OlympiadBench-COMP	physics_en	236		Acc
	SciKnowEval.L5	physics_problem_solving	200		LM
	SciEval	physics_knowledge_application	29		Acc
		physics_scientific_calculation	200		Acc
	UGPhysics	Electrodynamics	170		Acc
		Thermodynamics	200		Acc
		GeometricalOptics	54		Acc
		Relativity	200		Acc
		ClassicalElectromagnetism	200		Acc
		ClassicalMechanics	200		Acc
		WaveOptics	200		Acc
		QuantumMechanics	200		Acc
		TheoreticalMechanics	200		Acc
		AtomicPhysics	200		Acc
		SemiconductorPhysics	148		Acc
		Solid-StatePhysics	154		Acc
		StatisticalMechanics	200		Acc
	SuperGPQA	Physics	1482		Acc
Chemistry	GPQA	Chemistry	183	2158	Acc
	MMLU-Pro	chemistry	200		Acc
	SciBench	quan	41		Acc
		chemc	47		Acc
		atkins	121		Acc
		matter	57		Acc
	SciKnowEval.L5	chemical_procedure_generation	74		LM
		chemical_reagent_generation	125		LM
	SciEval	chemistry_knowledge_application	200		Acc
		chemistry_scientific_calculation	200		Acc
	SuperGPQA	Chemistry	910		Acc
Comp Sci	MMLU-Pro	computer science	200	415	Acc
	SciRIFF	Qasper	107		F1, LM
	SuperGPQA	Computer Science and Technology	108		Acc
Math	MMLU-Pro	math	200	2533	Acc
	SciBench	calc	52		Acc
		stat	92		Acc
		diff	55		Acc
	OlympiadBench-COMP	maths_en	674		Acc
	SuperGPQA	Mathematics	1460		Acc

Table 11: Domain-wise breakdown of SCIREAS tasks and instance counts (Part 1: Physics to Math).

Domain	Task Source	Subtask	Instances	Total	Metrics
Biology	GPQA	Biology	78	1911	Acc
	MMLU-Pro	biology	200		Acc
	LabBench	CloningScenarios	33		Acc
		ProtocolQA	108		Acc
		SeqQA	600		Acc
	SciKnowEval.L5	biological_procedure_generation	200		LM
		biological_reagent_generation	200		LM
	SciEval	biology_knowledge_application	200		Acc
		biology_scientific_calculation	200		Acc
	SuperGPQA	Biology	92		Acc
Medicine	MMLU-Pro	health	200	634	Acc
	SciRIFF	SciFact	184		F1, LM
		Evidence Inference	250		F1
Material Sci	SciKnowEval.L5	crystal_structure_and_composition	196	624	LM
		specified_band_gap_material_generation	200		LM
		property_and_usage_analysis	118		LM
	SuperGPQA	Materials Science and Engineering	110		Acc
Engineering	MMLU-Pro	engineering	200	2205	Acc
	SuperGPQA	Control Science and Engineering	77		Acc
		Information and Communication Engineering	156		Acc
		Electrical Engineering	234		Acc
		Chemical Engineering and Technology	226		Acc
		Power Engineering and Engineering Thermophysics	345		Acc
		Electronic Science and Technology	95		Acc
		Hydraulic Engineering	67		Acc
		Mechanics	456		Acc
		Mechanical Engineering	30		Acc
		Civil Engineering	93		Acc
		Optical Engineering	162		Acc
		Nuclear Science and Technology	30		Acc
		Instrument Science and Technology	12		Acc
		Systems Science	22		Acc

Table 12: Domain-wise breakdown of SCIREAS tasks and instance counts (Part 2: Biology to Engineering).

Domain	Task Source	Subtask/Subdomain	Instances	Total	Metrics
Physics	GPQA	Physics	8	388	Acc
	MMLU-Pro	physics	5		Acc
	SciBench	fund	1		Acc
		thermo	10		Acc
		class	8		Acc
	OlympiadBench-COMP	physics_en	25		Acc
	SciEval	physics_knowledge_application	1		Acc
		physics_scientific_calculation	1		Acc
	UGPhysics	Electrodynamics	17		Acc
		Thermodynamics	16		Acc
		GeometricalOptics	9		Acc
		Relativity	16		Acc
		ClassicalElectromagnetism	21		Acc
		ClassicalMechanics	17		Acc
		WaveOptics	16		Acc
		QuantumMechanics	17		Acc
		TheoreticalMechanics	13		Acc
		AtomicPhysics	13		Acc
		SemiconductorPhysics	13		Acc
		Solid-StatePhysics	13		Acc
		StatisticalMechanics	15		Acc
	SuperGPQA	Physics	133		Acc
Chemistry	GPQA	Chemistry	31	135	Acc
	MMLU-Pro	chemistry	3		Acc
	SciBench	quan	3		Acc
		chemc	2		Acc
		atkins	6		Acc
		matter	3		Acc
	SciEval	chemistry_knowledge_application	11		Acc
		chemistry_scientific_calculation	3		Acc
	SuperGPQA	Chemistry	73		Acc
Comp Sci	MMLU-Pro	computer science	6	21	Acc
	SuperGPQA	Computer Science and Technology	15		Acc
Math	MMLU-Pro	math	3	283	Acc
	SciBench	calc	2		Acc
		stat	2		Acc
		diff	3		Acc
	OlympiadBench-COMP	maths_en	92		Acc
	SuperGPQA	Mathematics	181		Acc

Table 13: Domain-wise breakdown of SCIREAS-PRO tasks and instance counts (Part 1: Physics to Math).

Domain	Task Source	Subtask	Instances	Total	Metrics
Biology	GPQA	Biology	2	123	Acc
	MMLU-Pro	biology	6		Acc
	LabBench	CloningScenarios	2		Acc
		ProtocolQA	10		Acc
		SeqQA	89		Acc
	SciEval	biology_knowledge_application	3		Acc
		biology_scientific_calculation	2		Acc
	SuperGPQA	Biology	9		Acc
Medicine	MMLU-Pro	health	5	5	Acc
Material Sci	SuperGPQA	Materials Science and Engineering	13	13	Acc
Engineering	MMLU-Pro	engineering	14	292	Acc
	SuperGPQA	Control Science and Engineering	7		Acc
		Information and Communication Engineering	15		Acc
		Electrical Engineering	32		Acc
		Chemical Engineering and Technology	43		Acc
		Power Engineering and Engineering Thermophysics	44		Acc
		Electronic Science and Technology	13		Acc
		Hydraulic Engineering	13		Acc
		Mechanics	54		Acc
		Mechanical Engineering	7		Acc
		Civil Engineering	18		Acc
		Optical Engineering	23		Acc
		Nuclear Science and Technology	3		Acc
		Instrument Science and Technology	2		Acc
		Systems Science	4		Acc

Table 14: Domain-wise breakdown of SCIREAS-PRO tasks and instance counts (Part 2: Biology to Engineering).